PRANJUL MISHRA
Warsaw, Poland • +48-508-491-119 • pranjulmishra228161@gmail.com
LinkedIn: linkedin.com/in/pranjul-mishra/ • GitHub: github.com/PM-0125

PROFILE

Machine Learning & Research Engineer with strong foundations in **algorithmic ML**, **RAG/NLP systems**, and **high-performance Python pipelines**. Experienced in taking research ideas from first principles to **production-grade code**, including packaging, testing, reproducibility, and CI/CD automation.Skilled in **structured retrieval (RDF/SPARQL)**, **LLM orchestration**, **parallel genomics algorithms**, and **end-to-end ML pipeline design**. Demonstrated ability to reason rigorously, learn rapidly, and deliver **robust, scalable, maintainable systems** under ambiguity.

CORE SKILLS

ML Engineering: PyTorch, TensorFlow, Scikit-learn, NLP, RAG, LLM prompting, Transformers, Knowledge Graphs
Python Engineering: OOP, Modular Design, Packaging, CLI Tools, pytest, mypy, ruff, black
Data & Infra: SQL, Pandas, NumPy, MLflow, DVC, GitHub Actions, Docker, Linux, DuckDB, QLever
Other: Parallel Computing, Algorithmic ML, Genomic Algorithms, Evaluation & Benchmarking
Strengths: First-principles reasoning, autonomous execution, rigorous engineering standards, rapid learning

---

SELECTED TECHNICAL ACHIEVEMENTS

- Implemented multiple **paper-to-code** algorithms, including SV phasing decision trees, retrieval reasoning logic, and statistical genomics methods.
- Built **3+ production-grade Python systems** with full CI, packaging, type-safety, tests, deterministic behaviour, and reproducibility.
- Designed multi-source retrieval pipelines integrating **RDF/SPARQL, DuckDB, OpenFDA APIs, UniProt, KEGG, Reactome, and LLM inference**.
- Delivered HPC pipelines scaling to **30–32 CPU cores**, processing full human genome datasets across **24 chromosomes**.
- Developed full-stack RAG system with **local LLM (Ollama), domain-aware templates, and multi-stage context selection**.

EXPERIENCE

Centre of New Technologies (CeNT) — University of Warsaw
Research Software Developer (Contract) · Jun 2025 – Present

- Refactored research prototypes into **production-grade Python libraries**, reducing integration issues and experimental failure modes by **~40%**.
- Designed reproducible ML setups (MLflow + DVC), enabling **disciplined experiment tracking** and saving **6–8 hours/week** in debugging and onboarding.
- Built deterministic CI workflows (pytest, mypy, ruff, GitHub Actions) enforcing **strict correctness guarantees** across pipelines.
- Developed benchmarkable evaluation harnesses for rapid exploration of algorithmic variants.

Marwadi University
Research Intern — Machine Learning · Jan 2023 – May 2023

- Developed automated ML-driven stock-trading strategies using Python + Pine Script, achieving **3× faster iteration cycles** via modular algorithmic design.
- Optimized model pipelines for **latency, stability, and scalability**, enabling high-frequency experimentation.

TSS Consultancy Pvt. Ltd.
Research Intern — NLP & Risk Analytics · Nov 2022 – Dec 2022

- Implemented components of a **large-scale financial risk-analysis pipeline** using NLP + ML for national banking workflows.
- Improved downstream reliability through **structured preprocessing, data validation**, and consistency checks.

PROJECTS

INFERMed — Clinical RAG System for Drug–Drug Interactions
Repo: https://github.com/PM-0125/INFERMed

- Built **INFERMed**, a full-stack RAG system combining RDF/SPARQL (PubChem via QLever), DuckDB-based clinical datasets (Parquet), and OpenFDA adverse event data.
- Designed modular retrieval clients for **UniProt, KEGG, Reactome, PubChem**, plus a PK/PD synthesis module aggregating pathways, enzymes, targets, and interaction evidence.
- Integrated a **local LLM (Ollama)** for structured generation, with domain-specific templates for Doctor / Patient / Pharma user modes.
- Served via a Streamlit front-end; engineered multi-stage context selection to improve reasoning quality. **Scale:** Processed **5M+ clinical records** across heterogeneous biomedical sources.

SvPhaser — Haplotype-aware SV Genotyper (Python Package, HPC)
Repo: https://github.com/SFGLab/SvPhaser

- Co-developed **SvPhaser**, an MIT-licensed haplotype-aware structural variant haplotyper, released as an importable Python package with CLI, documentation, and test suite.
- Implemented a deterministic **Δ-based decision tree algorithm** for phasing using HP-tagged BAM + VCF data.
- Parallelized across **24 chromosomes** using multiprocessing, scaling to **32-core machines** for genome-wide processing.
- Built modern packaging (pyproject.toml, src/ layout), pytest, type checks, and reproducible CI workflows. **Scale:** Processes **millions of long-read sequences** in genome analysis pipelines.

LOPHOS — Allele-specific Loops & Peaks Phasing Suite
Repo: https://github.com/SFGLab/lophos

- Built **LOPHOS**, a command-line suite for allele-specific phasing of peaks & chromatin loops from HiChIP BAMs.
- Implemented full statistical pipeline: read counting, **binomial tests**, FDR correction, and categorical bias classification (Maternal / Paternal / Balanced / Undetermined).
- Delivered a robust Python CLI with tests, pre-commit, type-safety, and GitHub Actions CI for automated validation. **Scale:** Handles **tens of GB** of HiChIP data with deterministic QC outputs.

EDUCATION

M.Sc., Computer Science & Information Systems (Artificial Intelligence)
Warsaw University of Technology · GPA: **4.63**
B.Tech., Computer Engineering (AI)
Marwadi University · GPA: 8.43

PUBLICATIONS & CERTIFICATIONS

- Academic Journal Article — University of Pitești (Erasmus+)
- Book Chapter — *Machine Intelligence & Big Data for Diabetes Management*
- Biomedical Journal — *Breast Cancer Survival Analysis*

Certifications:
NVIDIA Deep Learning Fundamentals · Microsoft Azure AI-900 · Oracle SQL · 10+ ML/DS Certificates

---

OTHER

Winner — Vodafone Idea National Innovation Marathon (2022)
Best Research Poster — AI in Bioinformatics (Anna University, 2022)
Languages: English (Fluent), Polish (Basic), Hindi (Native)